PENERAPAN K-*MEANS* DAN K-*MEDOIDS CLUSTERING* PADA DATA *INTERNET BANKING* DI BANK XYZ

APPLICATION OF K-MEANS AND K-MEDOIDS CLUSTERING ON INTERNET BANKING DATA IN XYZ BANK

Mediana Aryuni[1], E. Didik Madyatmadja[2], Eka Miranda[3]

School of Information Systems Binus University
Jl. K.H. Syahdan No. 9, Palmerah, Jakarta Barat 11480
[1]mediana.aryuni@binus.ac.id, [2]emadyatmadja@binus.edu, [3]ekamiranda@binus.ac.id[3]

Abstrak

Peningkatan jumlah pengguna *Internet Banking* berdampak pada semakin banyaknya data transaksi bank. Data tersebut bisa dimanfaatkan sebaik mungkin oleh bank. Salah satu caranya adalah menerapkan salah satu teknik *data mining*, yakni metode *clustering*. Tujuan penelitian ini adalah menerapkan metode *clustering* pada data transaksi *Internet Banking* untuk segmentasi *customer* sehingga dapat membantu pihak pemasaran bank. Metode penelitian mengacu pada metode *knowledge discovery*. Pada penelitian ini, diterapkan metode *unsupervised learning*, yakni *clustering* untuk mensegmentasi *customer*. Hasil penelitian menunjukkan bahwa nilai k yang paling optimal pada k-*means* adalah 3 dengan nilai *average within centroid distance* (W) sebesar 35.241. Sedangkan untuk k-*medoids*, nilai k yang paling optimal adalah 3 dengan nilai *average within centroid distance* (W) sebesar 88.849. Algoritme k-*means* memiliki performa yang lebih baik daripada k-*medoids*, baik dari sisi nilai *average within centroid distance* dan kompleksitas waktu.

Kata kunci: *clustering, data mining,* transaksi *internet banking*

Abstract

The increasing number of Internet Banking users has increased the number of bank transaction data. The data can be utilized as best as possible by the bank. The clustering method as one of data mining techniques can be applied to this data. The purpose of this research is to apply the clustering method on Internet Banking transaction data for customer segmentation in order to assist the marketing division of the bank. The research method refers to knowledge discovery methods (Han, Kamber, and Pei, 2012). In this study, an unsupervised learning method was applied, which was clustering to segment the customers. The results show that the most optimal k value in k-means is 3, where the average is within centroid distance (W) of 35.241. Meanwhile, the most optimal k value for k-medoids is 3, where the average is within centroid distance (W) of 88.849. The k-means algorithm has better performance compared to k-medoids observed from the average value within the centroid distance and the time complexity.

Keywords: *clustering, data mining, Internet Banking Transaction*

Tanggal Terima Naskah : 11 Desember 2017
Tanggal Persetujuan Naskah : 16 Januari 2018

1. PENDAHULUAN

Seiring dengan meningkatnya jumlah nasabah yang menggunakan *e-Banking* seperti yang dilansir oleh Kompas (2015), diikuti pula peningkatan dari segi volume dan frekuensi penggunaannya. Dari segi volume, penggunaan *e-Banking* meningkat dari Rp 4.441 triliun pada tahun 2012, menjadi Rp 5.495 triliun pada tahun 2013, dan menjadi Rp 6.447 triliun pada tahun 2014. Dari segi meningkatnya jumlah nasabah yang menggunakan *e-Banking*, pada tahun 2012 mencapai 3,79 miliar, pada 2013 meningkat menjadi 4,73 miliar, dan tahun 2014 mencapai 5,69 miliar.

Peningkatan jumlah pengguna *e-Banking* tentunya berdampak pada semakin banyaknya data transaksi bank. Data-data tersebut jika dibiarkan tentunya menjadi kurang berguna serta hanya memenuhi ruang penyimpanan data di bank. Oleh karena itu, diperlukan suatu cara untuk membuat agar data tersebut bisa dimanfaatkan sebaik mungkin oleh bank. Salah satu cara yang dapat digunakan, yaitu *data mining*.

Data mining dapat digunakan untuk melakukan segmentasi pasar dari nasabah pada sebuah bank [1]. Dari hasil *data mining* tersebut, dapat dianalisis hubungan antara profil nasabah dengan rasio transaksi *inflow* dan *outflow*, rasio saldo, beserta *channel* bank yang digunakan untuk melakukan transaksi. Pada penelitian tersebut *channel* bank yang diteliti meliputi kantor cabang bank, *e-Banking, m-Banking*, ATM, sistem, dan EDC. Terdapat dua fungsi utama dalam melakukan *data mining* di sektor perbankan, yakni untuk mempertahankan pelanggan (*customer retention*) serta deteksi penipuan (*fraud detection*) [2]. Penelitian [3] menerapkan algoritme K-Means Clustering untuk segmentasi *customer* yang efisien dalam strategi pada *targeted customer services*.

Penelitian ini akan membuat penerapan metode *clustering* pada data transaksi *Internet Banking* untuk segmentasi *customer* sehingga dapat membantu pihak pemasaran bank. Tujuan penelitian ini adalah menerapkan metode *clustering* untuk transaksi *Internet Banking* pada Bank XYZ agar diketahui metode yang tepat dalam mensegmentasi *customer*.

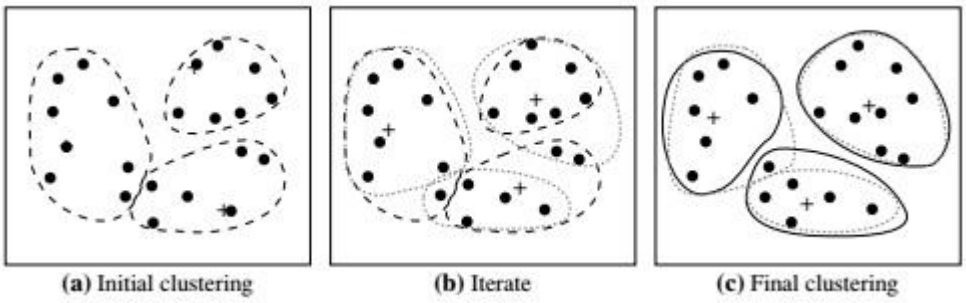
2. KONSEP DASAR

2.1 K-*Means Clustering*

Pada k-*means*, diberikan sebuah set data, D, dari sejumlah n objek, dan k, jumlah *cluster* yang akan dibentuk. Algoritma *partitioning* mengorganisasikan objek ke dalam partisi/*cluster* ($k \leq n$). Dari *cluster* yang telah terbentuk, dapat dihitung jaraknya untuk mengetahui objektivitas dari ketidaksamaan antar *cluster* [4].

Berikut adalah proses *clustering* dengan menggunakan metode k-*means* [4]:

a. Tentukan jumlah *cluster* (k) yang diinginkan.
b. Tentukan nilai *mean* yang akan menjadi pusat *cluster* awal.
c. Tetapkan setiap objek ke dalam *cluster* berdasarkan nilai *mean* objek yang paling mirip.
d. Perbarui nilai *mean* dari *cluster*, yaitu dengan menghitung nilai *mean* objek untuk setiap *cluster*.
e. Ulangi langkah 2-4 sampai tidak ada lagi perubahan pada nilai *mean* dari *cluster*. Ilustrasi dapat dilihat pada gambar 1.



Gambar 1. Proses *Clustering Partitioning Method* dengan menggunakan Metode K-*Means* [4]

2.2 K-*Medoids Clustering*

K-*means* berusaha meminimumkan nilai *total squared error*, sedangkan k-*medoids* meminimumkan *sum of dissimilarities* antara data di sebuah *cluster* dan memilih sebuah data di dalam *cluster* sebagai *center* (*medoids)* [5]. Gambar 2 menunjukkan algoritme K-*Medoids*.

Algorithm: *k*-medoids. PAM, a *k*-medoids algorithm for partitioning based on medoid or central objects.

Input:

- *k*: the number of clusters,
- *D*: a data set containing *n* objects.

Output: A set of *k* clusters.

Method:

(1) arbitrarily choose *k* objects in *D* as the initial representative objects or seeds;
(2) **repeat**
(3) assign each remaining object to the cluster with the nearest representative object;
(4) randomly select a nonrepresentative object, o_{random};
(5) compute the total cost, *S*, of swapping representative object, o_j, with o_{random};
(6) **if** $S < 0$ **then** swap o_j with o_{random} to form the new set of *k* representative objects;
(7) **until** no change;

Gambar 2. Algoritme K-*Medoids* [4]

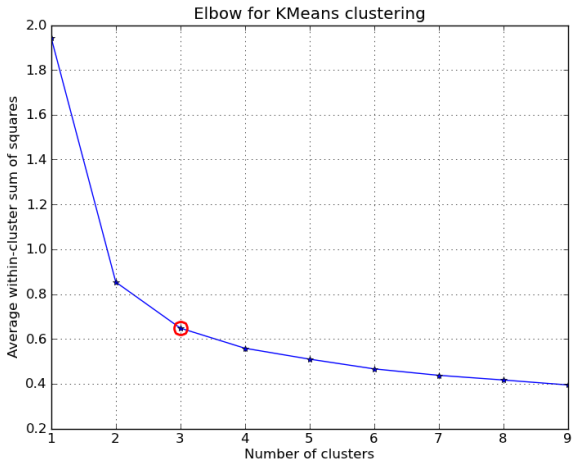
2.3 *Elbow Method* untuk Menentukan Nilai k (Jumlah *Cluster*) pada *Clustering*

Elbow method merupakan sebuah metode visual untuk menentukan nilai k yang optimal [6]. Caranya adalah dengan menentukan sebuah angka, k>0, *cluster* sejumlah k dibentuk dalam sebuah *data set* menggunakan algoritma *clustering* seperti *k-means*, selanjutnya dilakukan penghitungan *sum of within-cluster variance*, var(k). Langkah selanjutnya adalah pembuatan kurva antara nilai var dan k. Titik balik yang paling pertama (atau yang paling tajam) itulah yang merupakan jumlah yang "tepat" [4]. Contoh kurva *elbow method* dapat dilihat pada gambar 3.

Perhitungan *within-cluster sum of variance/error* (SSE atau var) untuk setiap jumlah k sesuai dengan persamaan 1 berikut ini:

$$\mathrm{SSE} = \sum \left(Y_i - \overline{Y}_i \right)^2 \quad \text{..... (1)}$$

dimana Y_i merupakan nilai dari setiap data pada *cluster* dan $\overline{Y_i}$ adalah nilai *centroid* dari *cluster*.



Gambar 3. Kurva *Elbow Method* [7]

3. METODE PENELITIAN

Penelitian ini menggunakan metodologi *knowledge discovery* [4] yang terdiri dari *Data cleaning, Data integration, Data selection, Data transformation,* dan *Data mining.*

4. HASIL DAN PEMBAHASAN

Data berasal dari beberapa tabel, sehingga dilakukan *data integration* untuk menghasilkan *dataset* yang siap diolah. *Dataset* yang digunakan merupakan gabungan dari enam tabel, yaitu transaksi *internet banking*, nasabah, produk, cabang, detail nasabah, dan saldo nasabah. Setelah dilakukan penggabungan dari keenam tabel tersebut, proses *data mining* menggunakan data tahun 2013 yang terdiri atas 2.964 transaksi. Tabel 1 menunjukkan 12 atribut yang dipilih untuk keperluan *data mining*. Pemilihan atribut dilakukan berdasarkan relevansi dengan tujuan yang ingin dicapai.

Tabel 1. Atribut-atribut yang dipilih untuk proses *data mining*

No.	Nama Atribut	Tipe Data Asli	Tipe Data setelah Transformasi	*Range* Data setelah Transformasi
1	*Gender*/jenis kelamin nasabah	Nominal	Numerik	*Male*: 0 *Female*: 1
2	*Birth date* nasabah	*Date*	Numerik	1 – 5 tahun = 0 6 – 10 tahun = 1 11 – 15 tahun = 2 16 – 20 tahun = 3 21 – 25 tahun = 4 26 – 30 tahun = 5 31 – 35 tahun = 6 36 – 40 tahun = 7 41 – 45 tahun = 8

Tabel 1. Atribut-atribut yang dipilih untuk proses *data mining* (Lanjutan)

No.	Nama Atribut	Tipe Data Asli	Tipe Data setelah Transformasi	*Range* Data setelah Transformasi
				46 – 50 tahun = 9 51 – 55 tahun = 10 56 – 60 tahun = 11 61 – 65 tahun = 12 66 – 70 tahun = 13 71 – 75 tahun = 14 76 – 80 tahun = 15 81 – 85 tahun = 16
3	Status perkawinan nasabah	Nominal	Numerik	Menikah = 0 Lajang/Belum Menikah = 1 Janda/Duda = 2
4	Kode profesi nasabah	Nominal	Numerik	Pelajar/ Mahasiswa = 0 Ibu Rumah Tangga = 1 Pegawai BUMN = 2 Pegawai Negeri = 3 Pegawai Swasta = 4 Wiraswasta = 5 ABRI/POLRI = 6 Profesional = 7 Pensiunan = 8 Pengacara, Notaris, Akuntan = 9 Guru, Dosen, dan sejenisnya = 10 Tidak diketahui (*unknown*) = 11 Lainnya = 12
5	Kode aplikasi	Nominal	Numerik	S (tabungan) = 0 D (giro/pinjaman rekening koran) = 1
6	Kode debit/kredit	Nominal	Numerik	D = 0 C = 1
7	Kota cabang	Nominal, misal Medan	Numerik	0-45
8	Provinsi cabang	Nominal, misal Sumatra Utara	Numerik	0-28
9	Nilai transaksi	Nominal	Numerik	1 – 500.000 = 0 500.001 – 2.500.000 = 1 2.500.001 – 7.500.000 = 2 7.500.001 – 15.000.000 = 3 15.000.001 – 50.000.000 = 4 50.000.001 – 100.000.000 = 5 100.000.001 – 250.000.000 = 6 250.000.001 – 500.000.000 = 7 > 500.000.000.000 = 8

Tabel 1. Atribut-atribut yang dipilih untuk proses *data mining* (Lanjutan)

No.	Nama Atribut	Tipe Data Asli	Tipe Data setelah Transformasi	*Range* Data setelah Transformasi
10	Saldo nasabah	Nominal	Numerik	< -5.000.000.000 = 0 -4.999.999.999 – -1.000.000.000 = 1 -999.999.999 – -500.000.000 = 2 -499.999.999 – 0 = 3 1 – 1.000.000 = 4 1.000.001 - 5.000.000 = 5 5.000.001 - 25.000.000 = 6 25.000.001 - 50.000.000 = 7 50.000.001 – 250.000.000 = 8 250.000.001 – 500.000.000 = 9 500.000.001 – 1.000.000.000 = 10 1.000.000.001 – 2.500.000.000 = 11 2.500.000.001 – 5.000.000.000 = 12 >5.000.000.000 = 13
11	Nama produk	Nominal	Numerik	Giro IDR Penduduk Perorangan = 0 Kewajiban Bolt = 1 Kewajiban Pokok ATM BCA = 2 XYZ *Payroll Group* Non Sinergi = 3 XYZ *Payroll Group* Sinergi = 4 XYZ *Savings* = 5 XYZ *Savings* (QQ *Savings*) = 6 XYZ *Savings* Lippo Homes = 7 XYZ *Savings* Plus = 8 XYZ *Savings* Promo = 9 XYZ *Savings* SDPDP = 10 PRK (Pinjaman Rekening Koran) IDR Penduduk Perorangan = 11 PRK (Pinjaman Rekening Koran) Perorangan Agunan Tunai = 12 Tabungan – Penduduk Passbook = 13
12	Tipe akun	Nominal	Numerik	Korporasi (CIBSGO) = 0 Perorangan (IBPERSON) = 1

Untuk membuat kurva dari *elbow method* ini, menggunakan *tools RapidMiner*. Total waktu eksekusi dari proses *elbow method* berturut-turut untuk k-*means* dan k-*medoids* adalah 3 detik dan 13 menit 44 detik. Hasil dari perhitungan proses tersebut dapat dilihat pada gambar 4.

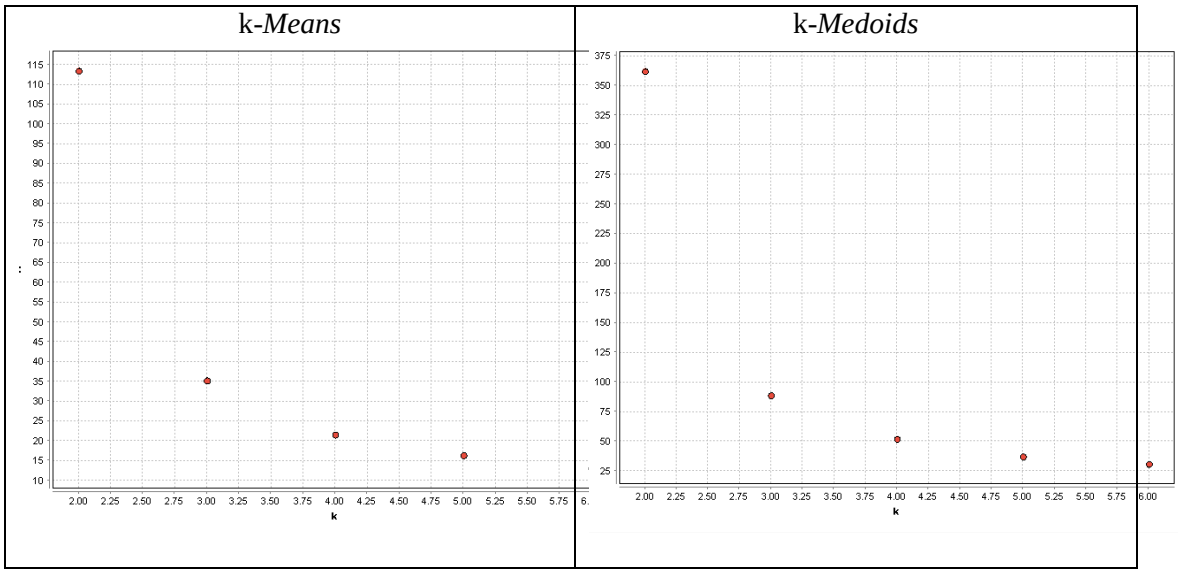
k-*Means* (Waktu: 3 detik)

k	W
2	113.436
3	35.241
4	21.569
5	16.343
6	13.091

k-*Medoids (*Waktu: 13 menit 44 detik)

k	W
2	362.056
3	88.849
4	52.066
5	37.217
6	30.958

Gambar 4. Perhitungan W (Rata-Rata Jarak Antara Tiap Data dengan *Centroid Cluster*) dan Waktu Komputasi untuk K=2 Sampai K=6



Gambar 5. Hasil Kurva *Elbow Method*

Berdasarkan hasil dari perhitungan *elbow method* pada gambar 5, dapat disimpulkan bahwa nilai k yang paling optimal pada k-*means* adalah 3 dengan nilai *average within centroid distance* (W) sebesar 35.241. Hal ini dikarenakan pada k=3 nilai W turun drastis, kemudian menjadi lebih stabil pada nilai k selanjutnya. Untuk k-*medoids*, nilai k yang paling optimal adalah 3 dengan nilai *average within centroid distance* (W) sebesar 88.849.

Dari hasil perbandingan nilai *average within centroid distance* pada gambar 5 menunjukkan bahwa algoritma k-*means* memiliki performa yang lebih baik dibandingkan algoritme k-*medoids*. Selain itu juga dari sisi kompleksitas waktu, k-*means* lebih baik daripada k-*medoids.*

5. KESIMPULAN

Penerapan metode *clustering* pada data *Internet Banking* menggunakan k-*means* dan k-*medoids* untuk melakukan segmentasi *customer*. Nilai k yang paling optimal pada k-*means* adalah 3 dengan nilai *average within centroid distance* (W) sebesar 35.241. Untuk k-*medoids*, nilai k yang paling optimal adalah 3 dengan nilai *average within centroid distance* (W) sebesar 88.849. Algoritma k-*means* memiliki performa yang lebih baik daripada k-

medoids baik dari sisi nilai *average within centroid distance* dan kompleksitas waktu.

REFERENSI

[1]. Sundjaja, Arta M., "*Analysis of Customer Segmentation in Bank XYZ Using Data Mining Technique*". *Asian Journal of Information Technology* 12(1): 39-44, ISSN: 1682-3915, 2013

[2]. Chitra, K. and Subashini, B. 2013. "Data *Mining Techniques and its Applications in Banking Sector*". *International Journal of Emerging Technology and Advanced Engineering*. ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 8.

[3]. Ezenkwu, C.P., Ozuomba, S., Kalu, C. 2015. "*Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Service". International Journal of Advanced Research in Artificial Intelligence*, Vol. 4, No.10.

[4]. Han, J., Kamber, M., Pei, J. 2012. Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann Publishers.

[5]. Mirkes, E.M. 2011. *K-means and K-medoids applet*. http://www.math.le.ac.uk/people/ag153/homepage/KmeansKmedoids/Kmeans_Kmedoids.html, University of Leicester.

[6]. Kodinariya, T. M., & Makwana, P. R. 2013. "*Review on Determining Number of Cluster in K-Means Clustering". International Journal of Advance Research in Computer Science and Management Studies*, *1*(6), 90-95. Retrieved from: https://www.researchgate.net/publication/303117595_Review_on_Determining_of_Cluster_in_K-means_Clustering

[7]. Stackoverflow. 2011. "*Calculating the percentage of variance measure for k-means".* https://stackoverflow.com/questions/6645895/calculating-the-percentage-of-variance-measure-for-k-means.